Perceptual integration of faces and voices depends on the interaction of emotional content and spatial frequency

Biological Psychology 123 (2017) 155–165

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Perceptual integration of faces and voices depends on the interaction of emotional content and spatial frequency

Jenny Kokinous [a,*], Alessandro Tavano [a,b], Sonja A. Kotz [c,d], Erich Schröger [a]

[a] Institute of Psychology, Leipzig University, Germany
[b] Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[c] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[d] Faculty of Psychology and Neuroscience, Dept. of Neuropsychology and Psychopharmacology, Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 11 October 2016
Accepted 11 December 2016
Available online 12 December 2016

Keywords:
Emotion
Audiovisual
Spatial frequency
EEG
Prediction

ABSTRACT

The role of spatial frequencies (SF) is highly debated in emotion perception, but previous work suggests the importance of low SFs for detecting emotion in faces. Furthermore, emotion perception essentially relies on the rapid integration of multimodal information from faces and voices. We used EEG to test the functional relevance of SFs in the integration of emotional and non-emotional audiovisual stimuli. While viewing dynamic face-voice pairs, participants were asked to identify auditory interjections, and the electroencephalogram (EEG) was recorded. Audiovisual integration was measured as auditory facilitation, indexed by the extent of the auditory N1 amplitude suppression in audiovisual compared to an auditory only condition. We found an interaction of SF filtering and emotion in the auditory response suppression. For neutral faces, larger N1 suppression ensued in the unfiltered and high SF conditions as compared to the low SF condition. Angry face perception led to a larger N1 suppression in the low SF condition. While the results for the neural faces indicate that perceptual quality in terms of SF content plays a major role in audiovisual integration, the results for angry faces suggest that early multisensory integration of emotional information favors low SF neural processing pathways, overruling the predictive value of the visual signal *per se*.

1. Introduction

Successful human communication depends on the accurate decoding of perceptual cues, such as a speaker's facial and vocal expression. An everyday visual communication scene contains a mixture of spatial frequencies (SFs) with low SFs conveying global and high SF conveying local stimulus features (e.g. De Cesarei & Codispoti, 2013). The role of SFs in visual emotion processing, and specifically in emotional face perception, is rather debated. Low SFs for example are processed primarily along the magnocellular visual pathway (e.g. Livingstone & Hubel, 1988), which is responsible for an initial fast and coarse visual analysis, and at the same time provides input to subcortical emotion areas such as the amygdalae, the pulvinar, and the superior colliculus (e.g. Vuilleumier, Armony, Driver, & Dolan, 2003). Therefore, a low SF advantage in the (early) perception of facial affect has been proposed at both behavioral (e.g. Kumar & Srinivasan, 2011 for happy expressions) and neural levels (e.g. Pourtois, Dan, Grandjean, Sander, & Vuilleumier, 2005; Vlamings, Goffaux, & Kemner, 2009; Vuilleumier et al., 2003). However, this assumption is still controversial (e.g. De Cesarei & Codispoti, 2013; Morawetz, Baudewig, Treue, & Dechent, 2011) and challenged by findings showing or suggesting that the use of SF information in emotion perception is flexible, for example, dependent on the task (e.g. Schyns & Oliva, 1999; Smith & Merlusca, 2014), the type of emotion (e.g. Fredrickson & Branigan, 2005; Srinivasan & Gupta, 2011), or the temporal unfolding of the visual content (De Cesarei, Mastria, & Codispoti, 2013; Holmes, Winston, & Eimer, 2005). A recent literature review concluded that a specialized role of low SFs in emotional processing cannot be favored (De Cesarei & Codispoti, 2013). However, the existing literature has preferentially focused on the link between SF and static emotion expressions and unisensory visual investigation, not considering that natural emotional communication comprises biological motion and is inherently multimodal. Therefore, the present study investigated the functional relevance of SFs in audiovisual integration of eco-

* Corresponding author at: Cognitive Incl. Biological Psychology, Institute of Psychology, Leipzig University, Neumarkt 9-19, 04109 Leipzig, Germany.
E-mail addresses: kokinous@uni-leipzig.de, jenny.kokinous@googlemail.com (J. Kokinous).

http://dx.doi.org/10.1016/j.biopsycho.2016.12.007
0301-0511/

logically valid, dynamic emotional and non-emotional faces and voices.

A crucial aspect of combining dynamic facial and vocal expressions is the time lag between the onset of the visual and the auditory signal (e.g. Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). It is assumed that the brain uses visual information to generate multisensory predictions about the quality and occurrence of the subsequent auditory signal, such as its location (spatial prediction), the time of auditory onset (temporal prediction) and the specific auditory features and informational content (formal prediction) (see e.g. Schwartze, Farrugia, & Kotz, 2013; Stekelenburg & Vroomen, 2007). Multimodal predictions typically facilitate audiovisual integration, resulting in a suppression and latency shortening of brain responses to audiovisual relative to unimodal stimuli occurring between 100 and 200 msec following auditory onset, representing the N1 and P2 component of the auditory event-related potential (ERP) (Baart, Stekelenburg, & Vroomen, 2014; Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003; Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014; Stekelenburg & Vroomen, 2007, 2012, 2015; van Wassenhove, Grant, & Poeppel, 2005). Although these neural mechanisms do not necessarily describe audiovisual integration in the sense of the formation of a newly integrated representation, they suggest that audiovisual interactions can occur very early in the processing stream in primary sensory brain areas (see also Baart et al., 2014; Calvert & Thesen, 2004; Koelewijn, Bronkhorst, & Theeuwes, 2010; Talsma, 2015). For non-emotional dynamic stimuli with natural visual-to-auditory delays, such as audiovisual speech or clapping hands, auditory N1 suppression has been shown to be modulated by temporal and spatial predictability (Stekelenburg & Vroomen, 2012; Vroomen & Stekelenburg, 2010) but not by formal predictability of the auditory stimulus, for example, it is insensitive to audiovisual incongruity (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). The suppression of the P2 response seems to be sensitive to the validity of formal predictions and to audiovisual congruity (e.g. Stekelenburg & Vroomen, 2007). Visual alterations caused by SF filtering also affect the informational content and formal predictive value of visual stimuli. Naturally, visual scenes are sometimes degraded, for example, in shortsightedness, peripheral vision, or fog. Thus, in varying the perceptual quality of visual stimuli in terms of SF content, the present study was suited to investigate formal multisensory predictions involved in audiovisual integration of dynamic faces and voices.

Emotional content encoded in visual information may allow for a stronger prediction of the auditory signal and may strengthen multisensory integration due to the saliency and evolutionary significance of emotions (Jessen & Kotz, 2013). Using dynamic emotional visual input such as moving faces or body movements, auditory N1 suppression has also been reported (Jessen & Kotz, 2011; Jessen, Obleser, & Kotz, 2012). Jessen & Kotz (2013) even showed that providing more visual information (a longer delay between visual and auditory onset) allowed for better prediction of the auditory input (larger N1 suppression) for emotional but not for neutral expressions. Additionally, we previously showed that emotional facial expressions lead to auditory N1 suppression regardless of the congruity with vocal information, while audiovisual congruity is required for the perceptual integration of neutral faces with voices (Kokinous, Kotz, Tavano, & Schröger, 2015).

The present ERP study examined the temporal properties of SF processing during audiovisual integration of ecologically valid, dynamic emotional and non-emotional faces and voices. SF filtering created either low- or high SF faces, removing the optimal SF range for face identification (mid to high, that is ∼5–20 cycles/face; Munhall et al., 2004; Näsänen, 1999) and reducing perceptual quality and formal predictive value of the visual stimuli while preserving temporal parameters and ecological validity. We aimed at exploring which SFs and perceptual strategies drive audiovisual integration of angry and neutral expressions, taking into account multisensory prediction. Unfiltered facial dynamics were used as a control condition. Incongruent audiovisual stimuli were additionally included (cf. Kokinous et al., 2015) to be able to unequivocally link the results to audiovisual integration. Based on previous findings, audiovisual integration was measured as latency modulation and amplitude suppression of the auditory N1 to audiovisual compared to auditory-only stimuli. In addition, we examined emotion categorization performance while processing audiovisual emotion expressions.

We specifically pursued the following questions:

i How does SF filtering affect auditory emotion identification?
ii How does SF filtering affect audiovisual integration? That is, are the latency and the amplitude (suppression) of the auditory N1 sensitive to the perceptual quality of the visual signal?
iii Does the effect of SF filtering on auditory N1 suppression interact with emotion, suggesting a difference in early audiovisual integration of emotional and non-emotional information?

2. Methods

2.1. Participants

Twenty-four healthy young adult volunteers took part in the EEG experiment; five of them were excluded from further analysis due to excessive alpha related artifacts. The remaining sample consisted of 19 participants (10 female; mean age = 23.7 years; SD = 4.8 years). Participants self-reported normal or corrected-to-normal vision and no hearing impairments. They gave written informed consent after the instruction of the experimental procedure, and received course credit or monetary reimbursement for participating in the study. Exclusion criteria included a history of brain injury, neurological disorder (e.g. stroke, epilepsy), any current treatment for mental illness or the intake of medication affecting the central nervous system. The experimental protocol adhered to the Declaration of Helsinki and the ethics guidelines of the German Association of Psychology.

2.2. Stimulus material and design

The original stimulus material had previously been developed and validated at the Max-Planck-Institute for Human Cognitive and Brain Sciences in Leipzig, Germany for research on multimodal affective processing (Ho, Schröger, & Kotz, 2014). Visual stimuli comprised a series of dynamic facial expressions of a 24-year young female speaker (portrait including hair and neck; mean duration = 1604 msec; range = 1000–2250 msec) expressing either a specific negative emotion (anger) or no emotion (neutral expression). Auditory stimulation consisted of a simultaneously spoken series of non-linguistic interjections (/ah/,/oh/) uttered by the same actress, expressing anger or no emotion (neutral). The delay between the onset of the visual and the onset of the auditory stimulus was variable due to a natural jitter in the individual recordings (mean = 765 msec). However, the V-A delay did not differ significantly between neutral (mean = 792,5 msec, SD = 312,4 msec) and angry (mean = 747,7 msec, SD = 95,8 msec) expressions ($t(34,4) = 0.75$, $p = 0.458$), ruling out the possibility that differences in the amount of available visual information at the time of voice onset between the emotion categories may explain potential emotion effects observed on audiovisual integration. The actress had been instructed to start each expression, emotional or non-emotional, with a neutral face to ensure that the emotion

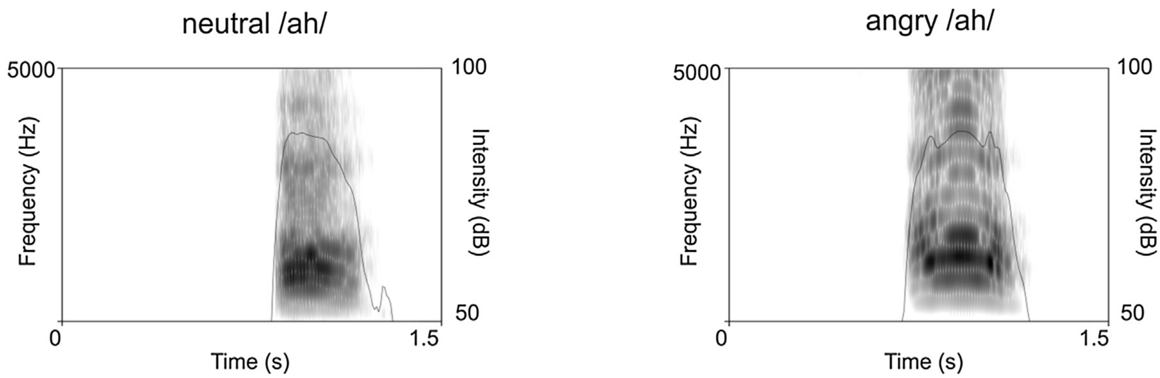

Fig. 1. Frequency spectrum and intensity contour of an example neutral/ah/(left) and an example angry/ah/(right).

evolved naturally in time. The intensity of neutral and angry interjections was controlled by normalizing the sound using the root mean square (RMS). No other manipulation was applied to the utterances to preserve their natural characteristics. Example frequency spectra of a neutral and an angry interjection used in the present experiment are plotted in Fig. 1, which illustrates that the intensity profile of neutral and angry stimuli is rather similar. However, they naturally differ in their frequency composition, that is, angry interjections contain a greater variability of frequencies than neutral interjections, which may serve as a physical cue for the emotion.

Several valence and arousal ratings and an emotion categorization study were performed on the stimulus material prior to the present experiment. A rating study (32 participants, 16 female), which used a two-dimensional valence and arousal rating space (Schubert, 1999) with manikins taken from Bradley and Lang (1994), confirmed that angry face-voice combinations were rated as more negative and more arousing than neutral face-voice pairs. An emotion classification study (40 additional participants, 20 female) investigated six basic emotions according to Ekman and Friesen (1976: anger, happiness, sadness, fear, disgust and neutral), and ensured that an expressed emotion was accurately and reliably recognized in the stimuli using only the face, only the voice, and audiovisual face-voice pairs, although performance was most accurate in latter condition (unbiased hit rate $H_u > 0.95$; see Wagner, 1993).

In the present experiment, a SF filtering technique was applied to the visual stimuli. Fifteen separately recorded videos per condition were selected. The originally colored videos were converted to gray-scale, and each frame of the video was individually filtered in the frequency domain using a Gaussian 2D filter. Image size was 432×432 pixels, including neck and background. A low-pass cutoff was set at 11 cycles/image (low-SF stimuli) and a high-pass cutoff was set at 41 cycles/image (high-SF stimuli), corresponding to the SF cutoffs used by Vuilleumier et al. (2003) on images that included only the face. First, Gaussian low-pass and high-pass filters were created in the frequency domain. Then, each frame underwent a Fast Fourier transformation and quadrant shift. The result was then multiplied by the respective filter, quadrants were shifted back, and data were translated again into the time dimension using the inverse of the Fast Fourier transform. Luminance was equated by subtracting the individual video minimum from each frame, and dividing by the individual video maximum minus the individual video minimum. To validate the videos, we investigated how SF filtering affects visual emotion identification in a behavioral experiment using the dynamic visual stimuli, and a sample that was independent of that from the subsequent electrophysiological recording prior to the present experiment. Specifically, we were interested whether there would be behavioral facilitation for low SF angry faces, following the hypothesis of a low SF advantage in emotion processing (e.g. Kumar & Srinivasan, 2011; Vuilleumier et al., 2003). In a three-alternative visual forced choice task, 18 participants (13 female; mean age = 25.5 years; SD = 4.8 years) identified the emotion expressed in the randomly presented, SF filtered (high-pass, low-pass or no filter) angry, happy and neutral dynamic facial expressions. Part A of the supplementary material shows and summarizes the findings of the behavioral experiment. Fig. S1 depicts behavioral performance in terms of unbiased hit rates (Wagner, 1993; Fig. S1a) and reaction times (RTs in msec, Fig. S1b) in the visual emotion categorization task. Globally, unbiased hit rates suggest a disadvantage for the identification of low SF faces whereas RTs indicate that both low and high SF faces are harder to recognize than unfiltered faces, and that the informative value gradually decreases from unfiltered to high SF to low SF faces. We also observed specific interactions of SF filter and emotional content. RTs were significantly longest for angry compared to happy and neutral faces in the low SF condition, and significantly shortest for happy faces in the high SF and the unfiltered condition. Thus, we did not find a low SF advantage and recognition speed was actually longest for low SF angry faces. However, accuracy was specifically enhanced for low SF angry faces while being equally high for all emotion categories in the unfiltered condition, and enhanced for happy faces in the high SF condition. This suggests that proposals of a low SF emotional advantage hold for dynamic faces. Possibly, the diagnostic cues for identifying angry dynamic facial expressions rely on low SF content, but the onset of such cues may be later for angry relative to happy and neutral dynamic expressions, which may explain the dissonant RT and accuracy findings observed for visual processing. For more detailed information, please see supplementary material. In the EEG experiment, audiovisual stimuli and solely angry and neutral expressions were used. We focused on the angry emotion category to be able to compare the present findings to those of previous studies (e.g. Ho et al., 2014; Kokinous et al., 2015). Anger is a highly relevant social emotion, given that it is a threat signal (Schupp et al., 2004) with negative valence and high arousal that demands behavioral adaptation from the observer (Frijda, 1986). It has been extensively studied and has inspired a large amount of experiments. Additionally, an auditory-only condition was required as a baseline to test for auditory suppression and audiovisual integration effects. Here, the interjections were presented while participants viewed a white fixation cross on a black computer screen.

2.3. Procedure

Participants sat comfortably in a sound-attenuated, electrically shielded and dimly lit chamber looking at a computer screen placed approximately 120 cm in front of them and holding a

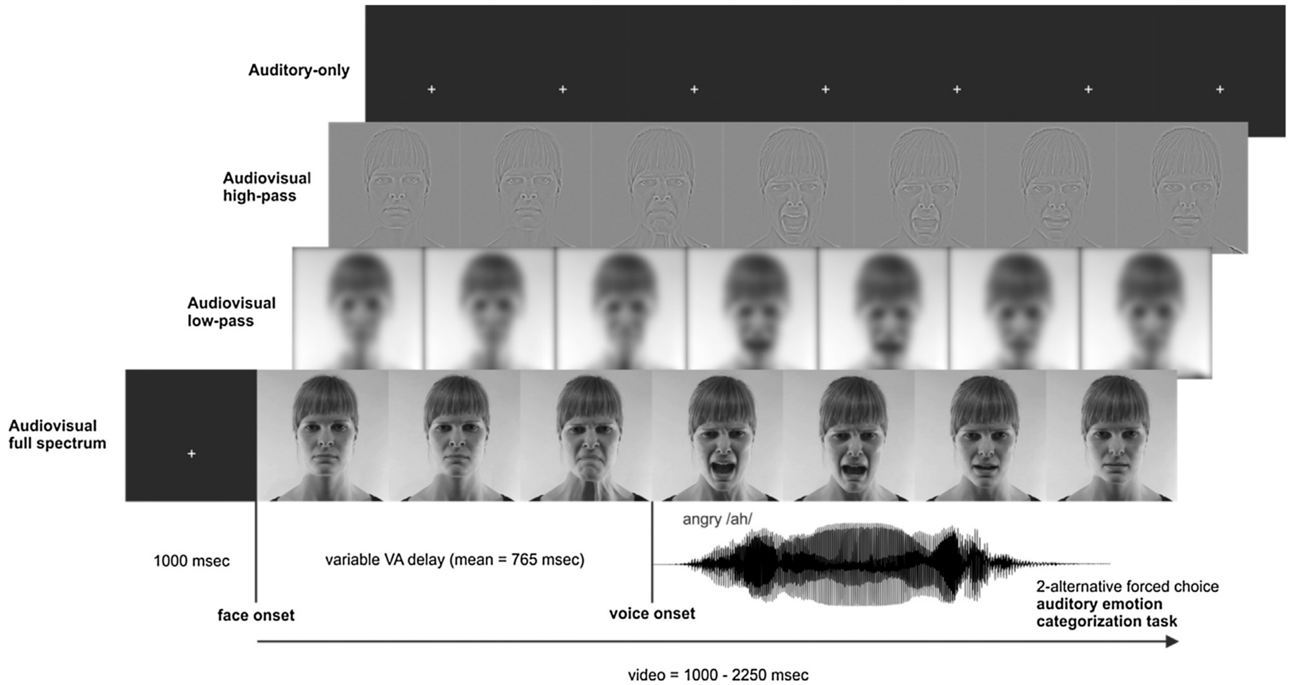

Fig. 2. Illustration of a trial, comprising the auditory-only condition and a schematic depiction of an example video in each SF condition (audiovisual full spectrum, high SF filtered, low SF filtered) with accompanying interjection (angry/ah/).

response device (Microsoft SideWinder Plug & Play Game Pad). Based on a previous study showing that audiovisual emotional congruity modulates perceptual integration in the N1 component (Kokinous et al., 2015), the EEG session included both congruent and incongruent audiovisual conditions and lasted approximately 83 min, not counting the breaks. The 120 trials per experimental condition (auditory-only neutral, auditory-only angry, audiovisual neutral unfiltered congruent, audiovisual neutral high-pass congruent, audiovisual neutral low-pass congruent, audiovisual angry unfiltered congruent, audiovisual angry high-pass congruent, audiovisual angry low-pass congruent, audiovisual neutral face unfiltered incongruent, audiovisual neutral face high-pass incongruent, audiovisual neutral face low-pass incongruent, audiovisual angry face unfiltered incongruent, audiovisual angry face high-pass incongruent and audiovisual angry face low-pass incongruent) added up to a total of 1680 trials. The experiment comprised 14 blocks (2 auditory, 12 audiovisual) of approx. 5.7 min duration. Trials were pseudo-randomized in a mixed design with a constant proportion of trials of each condition in each block. Based on previous studies (Ho et al., 2014; Kokinous et al., 2015), participants were asked to perform a two-alternative forced choice auditory emotion classification task ("Was the voice angry or not?" with "angry" and "not angry" as response options corresponding to a left and a right button on the response device). Sounds were presented binaurally via headphones (Sennheiser HD 25-1) at the same loudness level for all participants. Both speed and accuracy were emphasized and responses were given immediately during video presentation. At the beginning of a trial, a white fixation cross was presented on a black computer screen for 750 msec, followed by stimulus presentation (voice-only or audiovisual face-voice pair), and the participant's response. Responses were given with the thumbs of both hands and the response button assignment ("angry" left, "not angry" right vs. "not angry" left, "angry" right) and the order of experimental parts (auditory, audiovisual) was counterbalanced across participants. Before the start of the actual experiment, participants familiarized with the stimulus material by performing short training blocks (~2.7 min, 56 trials). The experiment was implemented using the Presentation software Version 15.0 (Neurobehavioral Systems, Inc.). Fig. 2 shows an example trial including a schematic depiction of a video in each SF condition with accompanying sound.

The electroencephalogram (EEG) was continuously recorded at a sampling rate of 500 Hz from 60 active Ag/AgCl electrodes mounted on an elastic cap (actiCAP; Brain Products GmbH, Munich, Germany) according to the international extended 10–20 system, and including the left and right mastoid. The signal was commonly referenced to a nose electrode, amplified by BrainVision Professional BrainAmp DC amplifiers and recorded with BrainVision Recorder software (Brain Products). Additional electrodes were placed on the outer canthi of both eyes (HEOG) as well as above and below the right eye (VEOG) to record eye-movements. The ground electrode was placed on the participant's forehead.

2.4. Data analysis

2.4.1. Behavioral data

Behavioral performance was assessed for each participant by computing reaction times (RTs in msec) and hit rates (in%) in the two-alternative auditory emotion categorization task.

We first collapsed the behavioral data across congruent and incongruent trials and computed repeated-measures ANOVAs with the factors voice emotion (neutral, angry) and visual context (auditory-only, audio + visual unfiltered, audio + visual high-pass filtered, audio + visual low-pass filtered). Comprising an auditory-only condition, this analysis was suited to investigate modality effects, e.g. potential performance benefits for audiovisual compared with auditory-only stimuli, and to examine the global influence of SF filtering on auditory emotion identification. Additionally, we were interested in whether voice identification performance would be facilitated for low SF angry faces across modalities, following assumptions of a low SF advantage in visual emotion processing.

In a second analysis, we explored whether SF filtering affects auditory emotion identification as a function of audiovisual congruity, since the experiment included congruent and incongruent audiovisual trials. Thus, to show potential incongruity effects and disentangle them from the effects of SF, we computed repeated-measures ANOVAs with the factors voice emotion (neutral, angry), audiovisual emotional congruity (congruent, incongruent) and SF filter (unfiltered, high-pass, low-pass) for both RTs and accuracy. The auditory-only condition was disregarded as it lacks the visual context.

2.4.2. Event-related potential (ERP) data

EEG data processing was performed with EEGLAB 12.0.2.5b (Delorme & Makeig, 2004) implemented in Matlab (Mathworks, Natick, MA). The electrophysiological data were filtered offline with a 0.5–30 Hz bandpass sinc FIR filter (Kaiser window, Kaiser beta 5.653, filter order 1812). Trials were averaged for each condition over a length of −100 to +700 msec in relation to the sound onset and the pre-onset time interval (−100 to 0 msec) was used as a baseline. Epochs containing samples exceeding amplitude changes of 75 μV were rejected. A minimum of 80 trials per averaged condition was ensured. The grand mean was calculated by averaging each of the conditions across participants. Statistical analyses were conducted on the brain responses time-locked to voice onset.

2.4.2.1. N1 latency. N1 peak latency was determined in each condition as the latency of the most negative peak of the individual ERP in a 70–140 msec time window that was chosen based on visual inspection of the grand average data. Subsequently, N1 latency was analyzed within a central region of interest (ROI: FCz, Cz, C1, C2, CPz) based on the N1's distribution considering the present data and the typical topography (e.g. Jessen & Kotz, 2011; Jessen et al., 2012). Scalp potential maps were computed between 70 and 140 msec (see Fig. 4).

Along the lines of the behavioral data analysis, we first averaged across congruent and incongruent conditions for the latency analysis and computed a repeated-measures ANOVA with the factors voice emotion (neutral, anger) and visual context (auditory-only, audio + visual unfiltered, audio + visual high-pass filtered, audio + visual low-pass filtered). This examined the influence of absent or present visual information of different perceptual quality on auditory processing speed (modality effects), and tested how SF filtering globally affects the time point of audiovisual integration.

In a second analysis, we specifically tested whether SF filtering affects N1 latency as a function of audiovisual congruity (incongruity effects). Therefore, we again calculated a repeated-measures ANOVA with the factors voice emotion (neutral, angry), audiovisual emotional congruity (congruent, incongruent) and SF filter (unfiltered, high-pass, low-pass), using the audiovisual conditions but disregarding the auditory-only condition as it lacks the visual context.

2.4.2.2. Auditory N1 suppression. We calculated the auditory N1 suppression effect by subtracting each audiovisual condition from the respective matched for emotion auditory-only condition (A − AV). Thus, specific auditory activity was removed, leaving only the contributions of the visual stimulus to auditory processing. The auditory suppression effect was analyzed within the described central ROI and the analysis window was determined around the mean of the individual condition peaks at electrode Cz (±15 msec). As the N1 suppression effect was clearly double-peaked at visual inspection, this resulted in a time window of 114–144 msec for the early N1 suppression effect and of 146–176 msec for the late N1 suppression effect (both N1 effects correspond to the latency of the right arm of the N100 component and to a time window commonly analyzed for N1 suppression effects; cf. e.g. Stekelenburg & Vroomen, 2007: 70–150 msec). Voltage distributions were created in the averaged time window of the early and late N1 suppression effect to illustrate the scalp topography of the auditory N1 suppression effect (see Fig. 5).

While the statistical analyses of the behavioral and N1 latency data focused on the emotion in the voice, the emotion factor in analysis of the N1 suppression effect was determined by the emotional content of the face (this distinction is relevant for incongruent trials). In the former analyses, it was mandatory to use voice emotion because we analyzed performance in an auditory task and compared processing in auditory-only and audiovisual situations (cf. to the approach in numerous other ERP studies on audiovisual integration, e.g. Stekelenburg & Vroomen, 2007, 2012; van Wassenhove et al., 2005). Previous findings, however, have also shown that facial information exert a larger influence on bimodal emotion processing than vocal information (e.g. Chen et al., 2015). Thus, there is a strong impact of the visual stimulus on audiovisual (emotion) integration and the visual stimulus drives early neuronal integration of dynamic audiovisual events, presumably via multisensory predictions (e.g. Jessen & Kotz, 2013, 2015; Jessen & Kotz, 2011; Kokinous et al., 2015; Stekelenburg & Vroomen, 2012). Therefore, the effects of manipulating visual stimulus parameters (SF) should be most pronounced when using face emotion for the analysis. Methodologically, this was implemented and legitimized using the difference waves and not the original ERPs for the statistical analysis, as described above. The different approaches to statistical analysis (using face vs. voice emotion) presumably tap into two different processing mechanisms, namely prediction driven by the visual cue and integration driven by the auditory stimulus.

Thus, to investigate the effects of SF filtering on N1 suppression, we submitted the mean amplitudes of the difference waves to a repeated-measures ANOVA with the factors time window (early N1, late N1), FACE emotion (neutral, angry), SF filter (unfiltered, high-pass, low-pass) and audiovisual emotional congruity (congruent, incongruent), also considering potential incongruity effects. We were again interested whether audiovisual integration would be facilitated for low SF angry faces (cross-modal low SF advantage). Modality effects could not be assessed using this analysis.

Visual inspection of the auditory ERPs (also Fig. 5) did not reveal a suppression of the P2 component in the audiovisual compared to the auditory-only condition for the angry emotion category. Thus, auditory P2 suppression effects were not analyzed. For effects and interactions that violated the assumption of sphericity, a Greenhouse-Geisser correction was applied. Appropriate follow-up ANOVAs and paired-samples *t*-tests were computed and corrected for multiple comparisons (p-value alpha-adjusted using the Bonferroni correction). Statistical analyses were conducted with the IBM SPSS Statistics software for Windows, Version 17 (IBM; Armonk, NY, USA).

To show that we could replicate our previous findings on the influence of incongruity on audiovisual emotion integration (cf. Kokinous et al., 2015), we computed voice emotion (neutral, angry) x visual context (auditory-only, audiovisual emotionally congruent, audiovisual emotionally incongruent) ANOVAs for N1 peak latency and peak amplitude on the present data, similarly to our previous work. For a detailed description of these analyses, the results and a figure (Fig. S2), see part B of the supplementary material.

3. Results

3.1. Behavioral results

Fig. 3 shows bar graphs depicting the two measures of behavioral performance derived from the auditory emotion discrimination task in the EEG experiment.

The voice conveyed emotion x visual context ANOVA for the RTs yielded a significant main effect of visual context (*F*(1,

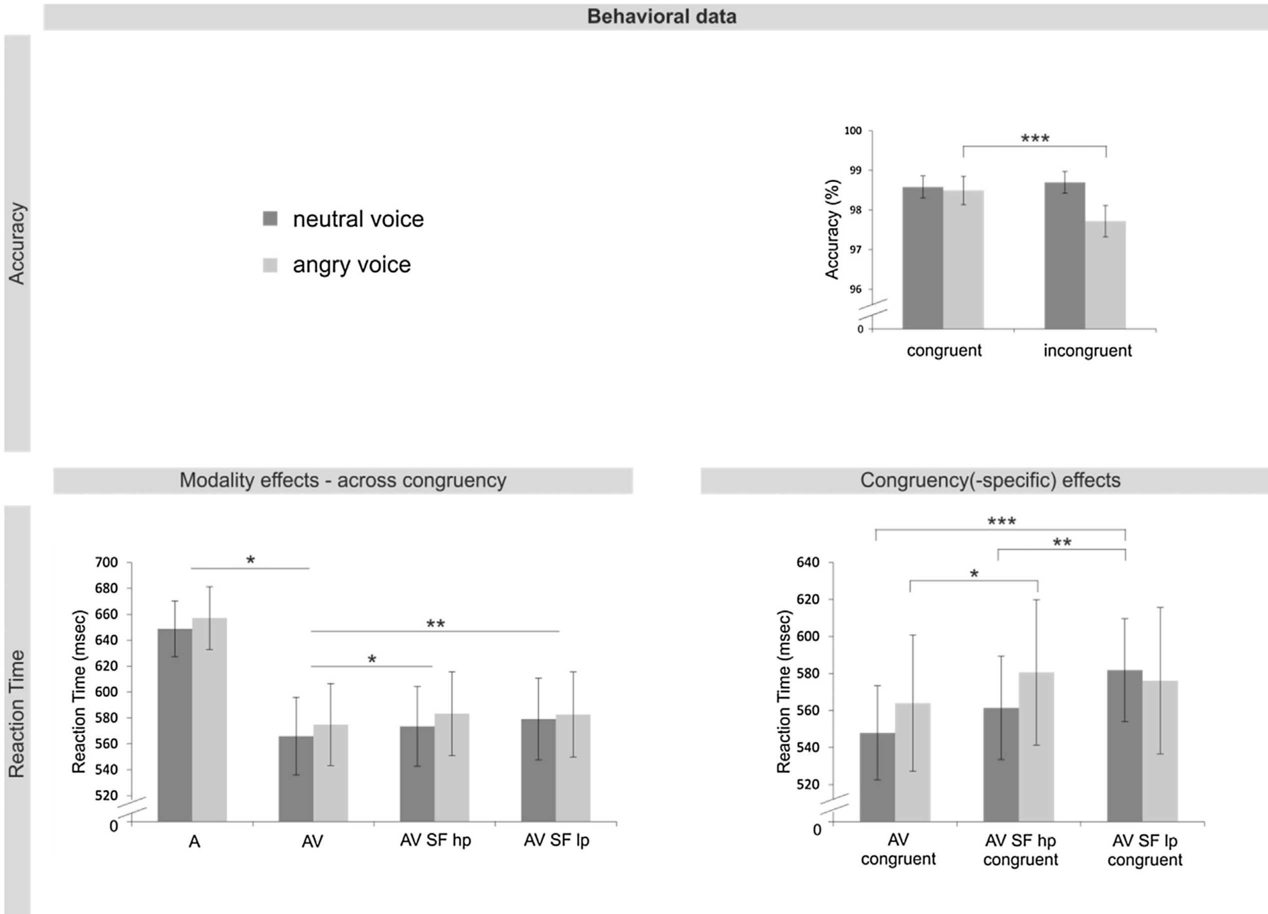

Fig. 3. Mean accuracy (top)/mean RTs (bottom) and standard errors in the two-alternative auditory forced choice task. Following the two-step approach of the statistical analysis, the effects of modality across congruent and incongruent trials (left) and the effects of audiovisual incongruity (right) are shown. Modality effects are shown only for the RTs because hit rates were generally at ceiling. For the incongruity effects, accuracy data was averaged across SFs to show the lower hit rates in the incongruent angry voice condition. RTs are depicted in the congruent condition and show the emotion x SF interaction that was also found for the ERPs, describing impairment of the identification of neutral voices paired with low SF faces that was not present for the identification of angry voices paired with low SF faces. Abbreviations: A, auditory-only; AV, audiovisual; SF, spatial frequency; hp, high-pass; lp, low-pass; ERP, event-related potential.

18.5) = 7.61, $p = 0.012$, $\eta^2 = 0.297$). Subsequent contrasts showed that participants responded significantly faster in the audiovisual unfiltered compared to the auditory-only condition (modality effect: $t(18) = 3.06$, $p = 0.041$). Additionally, RTs were significantly shorter in the audiovisual unfiltered condition than in both the audiovisual high SF ($t(18) = -2.97$, $p = 0.049$) and the audiovisual low SF ($t(18) = -3.62$, $p = 0.012$) conditions. Hit rates were generally at ceiling, that is, >98% in all conditions.

The voice emotion x audiovisual emotional congruity x SF filter ANOVA for the RTs yielded significant main effects of congruity ($F(1, 18) = 19.99$, $p < 0.001$, $\eta^2 = 0.526$) and SF filter ($F(2, 36) = 6.48$, $p = 0.004$, $\eta^2 = 0.265$) as well as a congruity x SF filter ($F(2, 36) = 7.54$, $p = 0.002$, $\eta^2 = 0.295$) and a voice emotion x congruity x SF filter ($F(2, 36) = 7.57$, $p = 0.002$, $\eta^2 = 0.296$) interaction. The latter was parsed by congruity and revealed congruity-specific interactions of emotion and SF. In the congruent condition, there was a significant voice emotion x SF filter interaction ($F(2, 36) = 5.28$, $p = 0.010$, $\eta^2 = 0.227$), and follow-up comparisons showed that for neutral voices, RTs were significantly longer in the low-pass compared with the high-pass ($t(18) = -3.48$, $p = 0.008$) and the full spectrum condition ($t(18) = -5.15$, $p < 0.001$). For angry voices, the low-pass condition did not differ significantly from the other two, but the full spectrum from the high-pass condition ($t(18) = -2.91$, $p = 0.028$) with RTs being generally longest in the high-pass condition. In the incongruent condition, the voice emotion x SF filter interaction was not significant. The accuracy analysis revealed significant main effects of voice emotion ($F(1, 18) = 7.31$, $p = 0.015$, $\eta^2 = 0.289$) and congruity ($F(1, 18) = 5.32$, $p = 0.033$, $\eta^2 = 0.228$) and a significant voice emotion x congruity interaction ($F(1,18) = 7.17$, $p = 0.015$, $\eta^2 = 0.285$). Parsing the interaction, pairwise comparisons across SF conditions showed that for angry (but not neutral) voices, accuracy was significantly lower in the incongruent compared to the congruent condition ($t(18) = 3.85$, $p = 0.001$).

3.2. ERP results

Fig. 4 shows the effects of modality and incongruity on N1 latency across SF conditions. Fig. 5 visualizes auditory N1 suppression across congruity trials in the three audiovisual filter conditions by means of ERPs and the auditory suppression effect (difference A-AV).

3.2.1. N1 latency

The voice emotion x visual context ANOVA yielded a significant main effect of emotion ($F(1, 18) = 120.55$, $p < 0.001$, $\eta^2 = 0.870$) with shorter N1 latencies for neutral (mean = 99 msec) compared to angry (mean = 111 msec) vocalizations. The main effect of visual context ($F(2,1, 37.3) = 72.1$, $p < 0.001$, $\eta^2 = 0.800$) was determined by significantly reduced latencies in all audiovisual conditions compared to the auditory-only condition across emotion and congruity categories (modality effect), as revealed by the follow-up contrasts (auditory-only = 116.2 msec, audiovi-

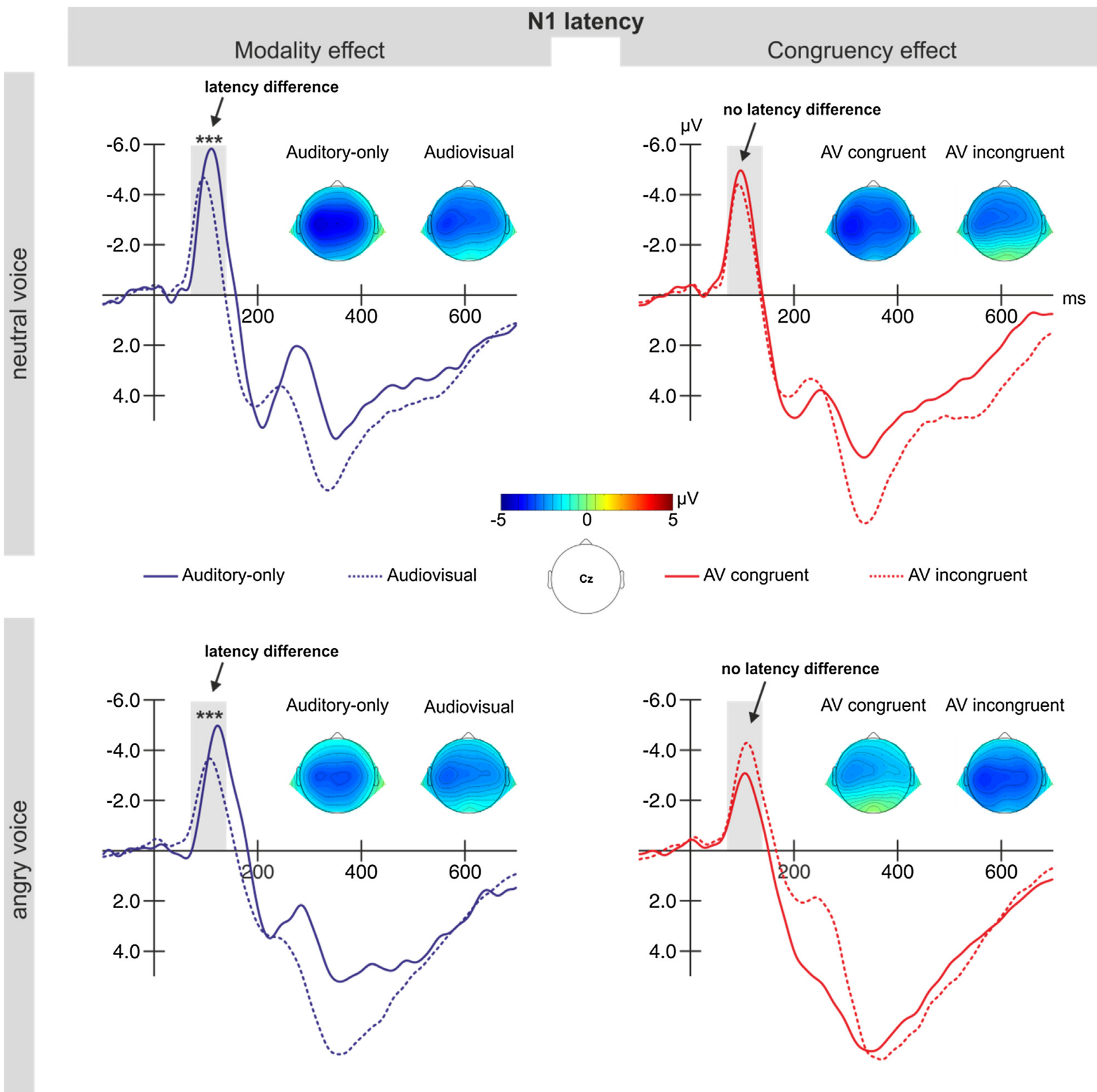

Fig. 4. N1 latency. **Left:** Auditory evoked potentials (AEPs) to voice onset at electrode Cz and corresponding N1 voltage distributions (70–140 msec) in the auditory-only and the averaged audiovisual conditions for neutral (top) and angry (bottom) voices. For visualization purposes, audiovisual ERPs were collapsed across SF filters because a first statistical analysis revealed filter-unspecific audiovisual N1 latency shortenings across congruency trials for both emotion categories (neutral and angry). **Right:** AEPs to voice onset at electrode Cz and N1 voltage maps in the time window used for the statistical analysis (70–140 msec) for congruent and incongruent audiovisual trials, plotted separately for neutral voices (top) and angry voices (bottom). ERPs were again collapsed across SFs because a second statistical analysis did not reveal significant differences in N1 latency between the SF conditions. Additionally, the figure illustrates that audiovisual congruency did not modulate N1 latency. Abbreviations: AV, audiovisual.

sual full spectrum = 100.2 msec: $t(18) = 10.41$, $p < 0.001$, audiovisual high-pass = 100.9 msec: $t(18) = 9.07$, $p < 0.001$, audiovisual low-pass = 101.5 msec: $t(18) = 13.52$, $p < 0.001$).

Except for a main effect of emotion ($F(1, 18) = 73.45$, $p < 0.001$, $\eta^2 = 0.803$), the voice emotion x audiovisual emotional congruity x SF filter ANOVA did not yield any significant results (no incongruity effects).

3.2.2. Auditory N1 suppression

The following results relate to the statistical analysis of the A – AV difference wave describing the auditory suppression effect. The time window x face emotion x SF filter x audiovisual congruity ANOVA yielded a significant main effect of emotion ($F(1, 18) = 8.42$, $p = 0.010$, $\eta^2 = 0.319$) indicating larger auditory N1 suppression for angry compared to neutral faces. Additionally, N1 suppression was larger for congruent compared with incongruent conditions, as revealed by the significant main effect of congruity ($F(1, 18) = 9.42$, $p = 0.007$, $\eta^2 = 0.344$). Follow-up comparisons for a significant interaction of time window and SF filter ($F(2, 36) = 3.38$, $p = 0.045$, $\eta^2 = 0.158$) did not yield any significant results. Thus, the influence of the SF filters on N1 suppression did not differ for early and late N1 suppression. The interaction of face emotion and SF filter was highly significant ($F(2, 36) = 12.20$, $p < 0.001$, $\eta^2 = 0.404$), suggesting that SF filtering differentially influenced auditory suppression of emotional and non-emotional stimuli across time windows and congruity conditions. The interaction was

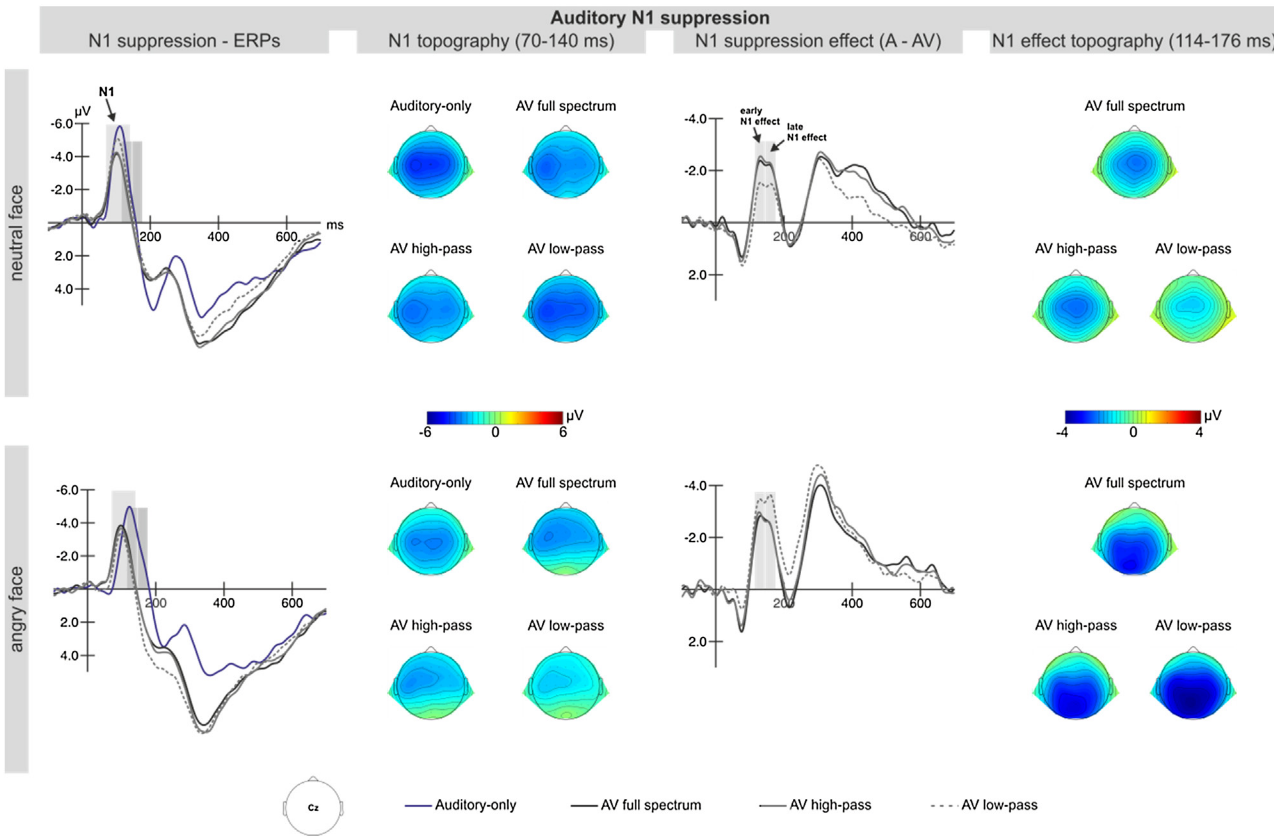

Fig. 5. Auditory N1 suppression, averaged across congruent and incongruent trials. **Left and middle left:** AEPs to voice onset and corresponding voltage distributions in the different audiovisual filter conditions, shown in relation to the auditory-only condition and plotted separately for neutral faces (top) and angry faces (bottom) at an exemplary electrode. Note that in accordance with the statistical analysis of the N1 suppression effect, audiovisual conditions are grouped by face emotion and not voice emotion whereas the auditory-only condition on the neutral face panel corresponds to the neutral voice condition and on the angry face panel to the angry voice condition (since there was no face). The light gray bars indicate the time window used for plotting the N1 topographies (70–140 msec). The dark grey bars indicate the time window used for the statistical analysis of the N1 suppression effect, demonstrating that the N1 suppression effects are largest on the right arm of the N100 ERP. The figure additionally illustrates the finding of an auditory emotion suppression effect with globally reduced amplitudes in the angry compared to the neutral conditions. **Middle right and right:** The auditory N1 suppression effect (measured as the difference of the auditory-only and the audiovisual ERPs) to voice onset and the corresponding voltage maps in the different audiovisual filter conditions, plotted separately for neutral (top) and angry (bottom) facial expressions at electrode Cz. The figure demonstrates that the low-pass condition selectively differs from the other audiovisual conditions in the N1 effect time windows, showing a pattern reversal for neutral and angry faces, and that N1 suppression was globally larger for angry compared to neutral faces. Analysis time windows are marked in gray and voltage maps are plotted in the combined time window of the early and the late N1 suppression effect (114–176 msec). Abbreviations: A, auditory-only; AV, audiovisual.

unpacked by face emotion in follow-up ANOVAs with the factor filter, which revealed significant filter effects for both neutral and angry faces ($F(2, 36) = 7.20$, $p = 0.002$, $\eta^2 = 0.286$ and $F(2, 36) = 5.37$, $p = 0.009$, $\eta^2 = 0.230$, respectively). Subsequent pairwise comparisons showed that in the neutral condition, the N1 suppression effect was significantly smaller in the low-pass compared with the full spectrum and the high-pass condition ($t(18) = -3.14$, $p = 0.017$ and $t(18) = 3.36$, $p = 0.010$, respectively). In terms of ERP amplitudes this means that the auditory N1 component was larger with low SF neutral faces compared with full and high SF neutral faces (and thus resembled more the amplitudes of the auditory-only condition). In the angry condition, the pattern reversed with significantly larger N1 suppression in the low-pass- compared to the full spectrum condition ($t(18) = 2.70$, $p = 0.044$). This means that low SF angry faces resulted in smaller auditory N1 amplitudes than full SF angry faces (and were thus more suppressed and less similar to the auditory-only condition).

4. Discussion

In the current study, SF filtering was applied to manipulate the perceptual quality and informational content of dynamic facial expressions. Expanding previous research, we assessed the relevance of SFs for multimodal processing of emotional and non-emotional faces and voices. Audiovisual integration was reflected as latency and amplitude facilitation of auditory brain responses in audiovisual compared with auditory-only conditions, due to the predictive nature of the stimuli. Auditory N1 suppression was determined as the main indicator for audiovisual integration. N1 latency and behavioral performance while processing audiovisual emotion expressions were additionally assessed.

Concerning the behavioral data from the auditory emotion discrimination task, a first analysis across congruity conditions showed an audiovisual behavioral benefit for the RTs with faster categorization in the audiovisual fully informative compared to the auditory-only context (modality effect), in accordance with numerous previous studies using comparable dynamic audiovisual emotion stimuli (Collignon et al., 2008; Föcker, Gondan, & Röder, 2011; Jessen & Kotz, 2011; Jessen et al., 2012; Klasen, Chen, & Mathiak, 2012; Kokinous et al., 2015; Massaro & Egan, 1996). The analysis also showed that RTs but not accuracy was sensitive to the influence of degraded visual signals on the discrimination of auditory input. That is, SF filtering (both low- and high-pass) prolonged the RTs and diminished the audiovisual benefit for the unfiltered audiovisual condition whereas accuracy was not reduced as a result of SF filtering. Thus, SF filtering globally caused a slowing of emotional voice processing. As was shown in the visual pilot experiment using the SF filtered videos, visual emotion identification speed was globally also impaired in both SF filter conditions compared to the unfiltered conditions. Thus, the visual information in these conditions cannot prime and facilitate auditory processing to the same extent as this would be possible for unfiltered visual identification. The system then must rely to a larger degree on the auditory information in order to solve the task, which explains the prolonged RTs for degraded visual stimuli. Additionally, while processing the auditory input, the degraded visual signal may lead to distraction and attentional reorienting towards the altered visual signal. Thus, across congruent and incongruent conditions, we did not find a behavioral advantage for the categorization of voices paired with low SF angry faces that would support proposals of a specialized role of low SFs in emotional processing. Importantly though, the second analysis that considered the congruity factor showed that audiovisual emotional incongruity influences behavioral performance (incongruity effects for RTs and accuracy). The hit rates showed a global impairment of the classification of angry voices by incongruity. Concerning the RTs, SF filtering did not affect voice categorization times for incongruent face-voice pairs – probably because SF effects were superimposed by the incongruity effects (attentional confounds) – whereas the SF content did make a difference for the categorization time of congruent expressions. There was a disadvantage, reflected in longer RTs, for the identification of neutral voices paired with low SF neutral faces that was not present for the identification of angry voices paired with low SF angry faces. Thus, addressing the question of a low SF advantage in emotional processing, low SF filtering left identification of angry voices unimpaired.

The electrophysiological data showed that visual SF filtering modulates auditory processing as early as ~100 msec starting from voice onset. This is consistent with other findings showing auditory suppression effects and audiovisual modulations of auditory processing at such relatively early perceptual processing stages (e.g. de Gelder, Böcker, Tuomainen, Hensen, & Vroomen, 1999; Jessen & Kotz, 2011; Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000; Stekelenburg & Vroomen, 2007). N1 latency was reduced in all audiovisual conditions compared to the auditory-only condition for both neutral and angry vocalizations (modality effect/audiovisual benefit) but not modulated by SF filtering or incongruity (but see behavioral findings). Thus, SF filtering did not affect the time point of audiovisual integration, neither globally nor as a function of audiovisual congruency. The global audiovisual latency reduction suggests early unspecific temporal facilitation in the integration of dynamic facial and vocal expressions. The absence of filter-specific latency modulations indicates that audiovisual integration is not based on different time points for low-pass, high-pass or unfiltered audiovisual expressions. Generally, speeded-up processing has been interpreted as a consequence of multisensory predictions in different bimodal situations (Jessen et al., 2012; van Wassenhove et al., 2005). Similar unspecific modality effects have been reported previously for the N1 under manipulations of emotional or phonetic audiovisual congruity (e.g. Kokinous et al., 2015; Stekelenburg & Vroomen, 2007). However, there have also been observations of congruity-specific audiovisual N1 latency shortenings (Knowland et al., 2014).

The analysis of the auditory N1 suppression effect showed that the ERPs elicited by audiovisual stimuli revealed the expected N1 amplitude reductions compared to the auditory only ERP response. There was no difference between early and late N1 suppression. N1 suppression was larger for angry compared to neutral facial expressions, indicating qualitative differences in auditory facilitation and audiovisual integration between neutral and emotional visual stimuli with stronger integration for the latter (see also Jessen & Kotz, 2013, 2015 for a literature review, and fMRI data), which cannot be due to differences in the amount of available visual information at the time of the voice onset between angry and neutral expressions. In agreement with assumptions that, due to the lack of ecological validity, integration and facilitation should not or to a lesser degree take place for incongruent audiovisual stimulation (e.g. Robins, Hunyadi, & Schultz, 2009), N1 suppression was globally reduced for incongruent face-voice pairs (incongruity effect). We did not find an interaction of emotion and congruity for N1 suppression, presumably since we investigated the impact of the visual stimulus on audiovisual integration using the emotional content of the face as a factor for the statistical analysis (inferred from suggestions by e.g. Baart et al., 2014; Chen et al., 2015; Jessen & Kotz, 2015, 2011 Jessen & Kotz, 2011; van Wassenhove et al., 2005). Further, using a statistical approach identical to that of our previous work on the incongruity effects on audiovisual integration (Kokinous et al., 2015), the present results fully align with our previous results (please see part B of the supplementary material). N1 suppression was modulated by SF in interaction with facial emotion, as there was a pattern reversal for neutral and angry faces in the low-pass condition. For neutral expressions, there was pronounced auditory suppression for fully informative and high-pass filtered faces, but smaller suppression for low-pass filtered facial expressions. In other words, the auditory N1 ERP was larger with low SF neutral faces compared with full and high SF neutral faces, and thus less facilitated with respect to the auditory-only condition. For angry facial expressions, the auditory N1 suppression was enhanced for the low-pass compared to the other filter conditions, i.e. low SF angry faces induced more facilitation of the N1 than full SF angry faces. Thus, audiovisual integration was sensitive to the perceptual quality of the visual signal and specifically susceptible to SF low-pass filtering, with different neural consequences for emotional and non-emotional stimuli. The behavioral data (specifically the RTs) support the interaction of emotion and SF found for N1 suppression, in that voice identification performance was altered for congruent audiovisual expressions with low SF neutral faces but not with low SF angry faces. Thus, the N1 suppression effects may be a neural index underlying the behavioral effects.

The results show that coarse low SF cues suffice to facilitate auditory processing and initiate audiovisual integration of dynamic emotional faces and voices, extending the debate from visual emotion research on the importance of SFs for emotion perception. Using a variety of methods and measures to investigate the relationship between emotion and SFs, it has been proposed that visual emotion processing is mainly dependent on low SFs in that emotional stimuli elicit behavioral, subcortical and cortical responses even when they contain only low SF cues (e.g. Pourtois et al., 2005; Vuilleumier et al., 2003; Winston, Vuilleumier, & Dolan, 2003). However, not all available data clearly support this hypothesis (De Cesarei & Codispoti, 2013), and there have also been divergent findings suggesting that visual emotion perception and emotion recognition is not selectively driven by low SFs (e.g. Holmes et al., 2005). The present finding that audiovisual integration of anger expressions prefers low SF information whereas the integration of non-emotional expressions is impaired for low SF information suggests the existence of distinct perceptual strategies and possibly the deployment of separate anatomical routes for the early integration of emotional and non-emotional facial expressions with vocalizations, comparable to suggestions of a dual-route model for visual processing of emotional and non-emotional faces (e.g. Vuilleumier et al., 2003). This is in line with views according to which the neural mechanisms of audiovisual integration differ for specific types of stimulation (see e.g. Baart et al., 2014 for speech-specific and non-speech integration), and suggests that audiovisual emotion integration may be qualitatively different from other kinds of audiovisual integration. In contrast to previous findings for non-emotional stimuli (Stekelenburg & Vroomen, 2007), N1 suppression here was already affected by the informational content of the visual signal (interaction of emotion and SF), thus occurring at a comparably earlier stage of audiovisual integration.

It should be noted that the specific role of N1 suppression in multimodal integration is still a matter of debate, and may represent multistage processes (e.g. Treille, Vilain, & Sato, 2014). As mentioned before, the effects found for low SF stimuli here therefore may not represent integration in the sense that a new mental representation in higher-order brain areas is formed, or that sensory inputs are integrated with existing cognitive schemata, but they may rather describe the modification of the original input signal in one modality by information from another modality (cf. multisensory interaction, Talsma, 2015). Additionally, it has been stated that suppression effects may generally not suffice to verify the occurrence of audiovisual (emotion) integration (Chen et al., 2015), but the global modulation of N1 suppression by audiovisual incongruity in our study suggests that it indeed does. Also, with respect to the low SF effects, it is possible that low SF filtering left intact more diagnostic cues for angry than for neutral faces, thus low SF filtering may simplify the extraction of the angry emotion. On the other hand, the emotional context of the angry face may generally facilitate audiovisual integration, overruling the informative value of the visual information per se.

Unlike in previous studies (Baart et al., 2014; Knowland et al., 2014; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005), we did not observe a clear auditory P2 suppression. Instead, in the angry face conditions P2 was reduced for auditory-only compared to audiovisual stimuli by visual inspection. For neutral expressions there was comparable auditory P2 suppression in all audiovisual conditions. We will not discuss the indications hereof as our research questions focused on the N1 but would like to point out that the discriminative effects of SF filtering on the N1 and the P2 are in line with other recent observations and suggestions that N1 and P2 can be dissociated concerning the influence of multisensory predictions and may reflect different stages of audiovisual integration (e.g. Baart et al., 2014; Ho et al., 2014; Jessen & Kotz, 2011; Stekelenburg & Vroomen, 2007, 2012; Vroomen & Stekelenburg, 2010).

Finally, to what extent the present findings are anger-specific remains to be investigated. Following previous studies (Ho et al., 2014; Jessen & Kotz, 2011; Kokinous et al., 2015), we selected anger as a strong representative for emotions due to its considerable social and evolutionary significance. We are aware that using one emotion category is highly selective, and that the generalization to other emotion categories needs to be undertaken. One limitation of the present study is that we cannot provide clear evidence of which emotion feature constitutes a specific effect, that is, whether they are due to the pure emotionality of the stimulus, the specific quality of the emotion, the physical parameters/visual features constituting the emotion, or the additional activation caused by the emotion. For this reason, future studies should also include other emotion categories. Another limitation of the present study was to use videos of only one identity and gender as it has been shown that there are gender differences in the processing of (facial) emotions (e.g. Bradley, Codispoti, Sabatinelli, & Lang, 2001; Forni-Santos & Osório, 2015; Wang et al., 2016). We decided not to vary speaker identity and gender for several reasons: for being able to compare the current findings to those of precursor studies (e.g. Ho et al., 2014; Kokinous et al., 2015), due to experimental time limitations as the experiment was rather long (~83 min, excluding preparation time), and due to the fact that we wanted to avoid "noise" with respect to the filter/emotion effects possibly being introduced by identity and/or gender main (or interaction) effects. Nevertheless, future studies should consider all gender displays.

5. Conclusion

SF filtering was applied to dynamic facial expressions to investigate the role of SF for audiovisual emotion integration. SF filtering differentially affected auditory N1 suppression, an indicator for audiovisual integration, of emotional and non-emotional stimuli. We provide evidence for impaired audiovisual integration of SF low-pass filtered neutral faces with voices (less N1 suppression), but enhanced integration of low-pass filtered angry faces with voices (more N1 suppression). We conclude 1) that audiovisual integration is sensitive to the perceptual quality of the visual signal in terms of SF content, and 2) that there is a difference in early perceptual integration of emotional and non-emotional audiovisual expressions, whereby audiovisual emotion integration relies primarily on low SFs. The present study adds to the literature by providing information about the relation of SF and audiovisual integration. By using dynamic emotional expressions it allows the investigation of the temporal properties of SF processing and ensures a more ecologically valid understanding of the perception of multimodal emotions.

Acknowledgement

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the research training group 1182 "Function of Attention in Cognition" (scholarship to J.K.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biopsycho.2016.12.007.

References

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115–121. http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *The European Journal of Neuroscience*, *20*(8), 2225–2234. http://dx.doi.org/10.1111/j.1460-9568.2004.03670.x

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Bradley, M. M., Codispoti, M., Sabatinelli, D., & Lang, P. J. (2001). Emotion and motivation II: Sex differences in picture processing. *Emotion*, *1*(3), 300–319. http://dx.doi.org/10.1037/1528-3542.1.3.300

Calvert, G. A., & Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, *98*(1–3), 191–205. http://dx.doi.org/10.1016/j.jphysparis.2004.03.018

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. http://dx.doi.org/10.1371/journal.pcbi.1000436

Chen, X., Pan, Z., Wang, P., Yang, X., Liu, P., You, X., et al. (2015). The integration of facial and vocal cues during emotional change perception: EEG markers. *Social Cognitive and Affective Neuroscience*, nsv083. http://dx.doi.org/10.1093/scan/nsv083

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., et al. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135. http://dx.doi.org/10.1016/j.brainres.2008.04.023

De Cesarei, A., & Codispoti, M. (2013). Spatial frequencies and emotional perception. *Reviews in the Neurosciences*, *24*(1), 89–104. http://dx.doi.org/10.1515/revneuro-2012-0053

De Cesarei, A., Mastria, S., & Codispoti, M. (2013). Early spatial frequency processing of natural images: an ERP study. *Public Library of Science*, *8*(5), e65103. http://dx.doi.org/10.1371/journal.pone.0065103

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. http://dx.doi.org/10.1016/j.jneumeth.2003.10.009

de Gelder, B., Böcker, K. B., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neuroscience Letters*, *260*(2), 133–136.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto. CA: Consulting Psychologists Press.

Föcker, J., Gondan, M., & Röder, B. (2011). Preattentive processing of audio-visual emotional signals. *Acta Psychologica*, *137*(1), 36–47. http://dx.doi.org/10.1016/j.actpsy.2011.02.004

Forni-Santos, L., & Osório, F. L. (2015). Influence of gender in the recognition of basic facial expressions: a critical literature review. *World Journal of Psychiatry*, *5*(3), 342–351. http://dx.doi.org/10.5498/wjp.v5.i3.342

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion*, *19*(3), 313–332. http://dx.doi.org/10.1080/02699930441000238

Frijda, N. H. (1986). *The emotions*. Cambridge University Press.

Ho, H. T., Schröger, E., & Kotz, S. A. (2014). Selective attention modulates early human evoked potentials during emotional face–voice processing. *Journal of Cognitive Neuroscience*, 1–21. http://dx.doi.org/10.1162/jocn_a_00734

Holmes, A., Winston, J. S., & Eimer, M. (2005). The role of spatial frequency information for ERP components sensitive to faces and emotional facial expression. *Brain Research Cognitive Brain Research*, *25*(2), 508–520. http://dx.doi.org/10.1016/j.cogbrainres.2005.08.003

Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage*, *58*(2), 665–674. http://dx.doi.org/10.1016/j.neuroimage.2011.06.035

Jessen, S., & Kotz, S. A. (2013). On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, *7*, 369. http://dx.doi.org/10.3389/fnhum.2013.00369

Jessen, S., & Kotz, S. A. (2015). Affect differentially modulates brain activation in uni- and multisensory body-voice perception. *Neuropsychologia*, *66*, 134–143. http://dx.doi.org/10.1016/j.neuropsychologia.2014.10.038

Jessen, S., Obleser, J., & Kotz, S. A. (2012). How bodies and voices interact in early emotion perception. *Public Library Of Science*, *7*(4), e36070. http://dx.doi.org/10.1371/journal.pone.0036070

Klasen, M., Chen, Y.-H., & Mathiak, K. (2012). Multisensory emotions: perception, combination and underlying neural processes. *Reviews in the Neurosciences*, *23*(4), 381–392. http://dx.doi.org/10.1515/revneuro-2012-0040

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research Cognitive Brain Research*, *18*(1), 65–75.

Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. (2014). Audio-visual speech perception: a developmental ERP investigation. *Developmental Science*, *17*(1), 110–124. http://dx.doi.org/10.1111/desc.12098

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384. http://dx.doi.org/10.1016/j.actpsy.2010.03.010

Kokinous, J., Kotz, S. A., Tavano, A., & Schröger, E. (2015). The role of emotion in dynamic audiovisual integration of faces and voices. *Social Cognitive and Affective Neuroscience*, *10*(5), 713–720. http://dx.doi.org/10.1093/scan/nsu105

Kumar, D., & Srinivasan, N. (2011). Emotion perception is mediated by spatial frequency content. *Emotion (Washington, D.C.)*, *11*(5), 1144–1151. http://dx.doi.org/10.1037/a0025453

Livingstone, M., & Hubel, D. (1988). *Segregation of form, color, movement, and depth: anatomy, physiology, and perception*. [Retrieved from http://papers.cumincad.org/cgi-bin/works/_id=ecaade2013/Show?4744]

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215–221. http://dx.doi.org/10.3758/BF03212421

Morawetz, C., Baudewig, J., Treue, S., & Dechent, P. (2011). Effects of spatial frequency and location of fearful faces on human amygdala activity. *Brain Research*, *1371*, 87–99. http://dx.doi.org/10.1016/j.brainres.2010.10.110

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics*, *66*(4), 574–583.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, *39*(23), 3824–3833.

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport*, *11*(6), 1329–1333.

Pourtois, G., Dan, E. S., Grandjean, D., Sander, D., & Vuilleumier, P. (2005). Enhanced extrastriate visual response to bandpass spatial frequency filtered fearful faces: time course and topographic evoked-potentials mapping. *Human Brain Mapping*, *26*(1), 65–79. http://dx.doi.org/10.1002/hbm.20130

Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain and Cognition*, *69*(2), 269–278. http://dx.doi.org/10.1016/j.bandc.2008.08.007

Schubert, E. (1999). Measuring emotion continuously: validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154–165. http://dx.doi.org/10.1080/00049539908255353

Schupp, H. T., Ohman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: an ERP analysis. *Emotion (Washington, D.C.)*, *4*(2), 189–200. http://dx.doi.org/10.1037/1528-3542.4.2.189

Schwartze, M., Farrugia, N., & Kotz, S. A. (2013). Dissociation of formal and temporal predictability in early auditory evoked potentials. *Neuropsychologia*, *51*(2), 320–325. http://dx.doi.org/10.1016/j.neuropsychologia.2012.09.037

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, *69*(3), 243–265.

Smith, M. L., & Merlusca, C. (2014). How task shapes the use of information during facial expression categorizations. *Emotion (Washington, D.C.)*, *14*(3), 478–487. http://dx.doi.org/10.1037/a0035588

Srinivasan, N., & Gupta, R. (2011). Rapid communication: global-local processing affects recognition of distractor emotional faces. *Quarterly Journal of Experimental Psychology (2006)*, *64*(3), 425–433. http://dx.doi.org/10.1080/17470218.2011.552981

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973. http://dx.doi.org/10.1162/jocn.2007.19.12.1964

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*, 26. http://dx.doi.org/10.3389/fnint.2012.00026

Stekelenburg, J. J., & Vroomen, J. (2015). Predictive coding of visual-auditory and motor-auditory events: an electrophysiological study. *Brain Research*, *1626*, 88–96. http://dx.doi.org/10.1016/j.brainres.2015.01.036

Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, *9*, 19. http://dx.doi.org/10.3389/fnint.2015.00019

Treille, A., Vilain, C., & Sato, M. (2014). The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. *Frontiers in Psychology*, *5*, 420. http://dx.doi.org/10.3389/fpsyg.2014.00420

Vlamings, P. H. J. M., Goffaux, V., & Kemner, C. (2009). Is the early modulation of brain activity by fearful facial expressions primarily mediated by coarse low spatial frequency information? *Journal of Vision*, *9*(5), 1–13. http://dx.doi.org/10.1167/9.5.12 [12]

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596. http://dx.doi.org/10.1162/jocn.2009.21308

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience*, *6*(6), 624–631. http://dx.doi.org/10.1038/nn1057

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3–28. http://dx.doi.org/10.1007/BF00987006

Wang, S., Li, W., Lv, B., Chen, X., Liu, Y., & Jiang, Z. (2016). ERP comparison study of face gender and expression processing in unattended condition. *Neuroscience Letters*, *618*, 39–44. http://dx.doi.org/10.1016/j.neulet.2016.02.039

Winston, J. S., Vuilleumier, P., & Dolan, R. J. (2003). Effects of low-spatial frequency components of fearful faces on fusiform cortex activity. *Current Biology: CB*, *13*(20), 1824–1829.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186. http://dx.doi.org/10.1073/pnas.0408949102